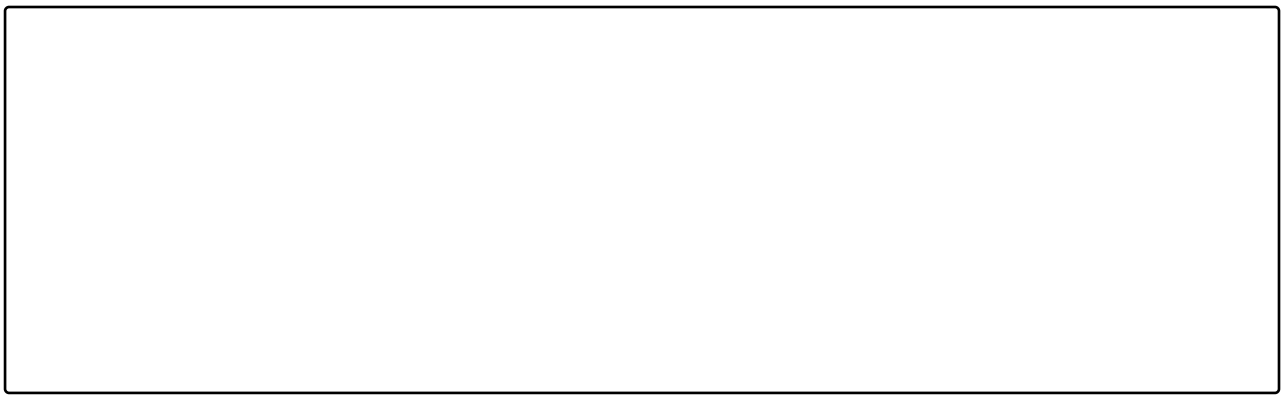
4 Programming Part 1: PCA, Clustering, and SVM [20 points]

This part of the written assignment should be done after you complete the first part of the programming:

1. **[3 pts]** After completing K-Means, attach the scatterplot of the 2D PCA Projection with K-Means Clusters. The histograms show how many data points from each year belong in a specific cluster. One of the years should stand out. Report that year, and make a prediction on what might have happened that year to cause this phenomenon.

2. **[3 pts]** Try out 4 different kernels and for each kernel, fit an SVM on the same dataset. Report the 4 accuracies (report the hyper-parameters you chose too). Which kernel performed the best? Which kernel performed the worst? Did you expect that?

3. **[2 pts]** In the last cell, we fit an SVM on the labeled stock data using K-means where $K = 2$. Is the accuracy higher or lower than what you had before? Is that surprising? Why or why not?



5 Programming Part 2: Kaggle Competition [30 points]

For the 2nd part of the programming, you will take what you learned throughout this course and apply it to a real-world scneario! Kaggle is one of the more well-known data science/machine learning competitions, so this is a start if you ever want to participate in the future!

5.1 Beginning Steps

1. Create an account on kaggle.com
2. Enter the competition through this link
3. Every time you want to go back to the website, you can directly enter through here.

Get used to the page! Specifically, go through the 'Overview' and 'Data' page. You will be given both the training and testing set, where the training set has the labels. You will submit your predictions for the test set, and 80% of the test set will count to the public leaderboard (which is public). After this assignment closes, the other 20% will be used for the final leaderboard!

This is pretty new to you all, so feel free to ask on Piazza if anything seems unclear!

5.2 The Assignment

For this assignment, you will fully analyze the data, and create a machine learning model that creates the best predictions. More specifically, you should submit a jupyter notebook that:

1. Analyzes the data before fitting it with a model
2. Cleans the data/adds other features to improve the data
3. Fits a model to see what happens
4. Tune hyper-parameters to obtain better results
5. Implement some bagging/boosting to improve results

The baseline model is also provided to you so you know how to get started. You could use Google Collab, your own machine, the online interpreter on Kaggle etc., as long as you submit the notebook to gradescope and the predictions to Kaggle.

The rubric is provided at the end of this assignment. In other words, you can earn full credit even if you aren't at the top of the leaderboard; You just need to put enough effort into this assignment.

Enter your name on the Kaggle leaderboard here:

5.3 Getting Help

For this part of the assignment only, **You are allowed to directly lookup/implement models online. In fact, you are encouraged to search from other Kaggle notebooks for inspiration. However, you should still cite your sources and directly copying someone else's code is still plagiarism**. Most of the Kaggle community is collaborative, where people publish notebooks to guide people to implement a certain model, and you are allowed to find help online. In fact, this contains some of the best notebooks from this competition, and this contains insight from several people on advice to tackle this dataset.

If you use a specific strategy from online, **CITE THEM IN YOUR CODE!** We will know if you directly copied someone's code without understanding what is going on.here are some pretty

5.4 Extra Credit

As an incentive to reach the top of the leaderboard, the **top 5 submissions** will receive special 10-315 stickers! (You will also catch our attention if you ever apply to be an ML TA/you can put it on your resume)! Again, you can be last on the leaderboard and STILL get full credit if you put in the effort mentioned in the last page.

5.5 Rubric

Criteria	5 Points	2.5 Points	0 Points
Exploratory Data Analysis	Performs in depth data analysis, shows different graphs and understanding of relationships between data	Shows different graphs for the purpose of showing graphs	No effort to analyze data
Data Cleaning	Does through cleaning that improves data quality	Uses traditional scaling methods mentioned in class	Does little to no data cleaning
Model Fitting	Uses advanced models and tried different hyper-parameters	Uses traditional models with some attempt to improve	Directly uses model from sklearn/torch without tuning
Bagging/Boosting	Combines several methods together appropriately	Does naive bagging such as averaging/taking max (unless explicitly states good reason)	No attempt at bagging/boosting, or directly uses algorithm such as XGBoost with no effort
Creativity	Shows originality and creative thoughts in this assignment	Shows average amount of work	Shows minimal effort in assignment
Submission	Submits guesses on Kaggle, writes Kaggle name in assignment, submits notebook to Gradescope	Missing 1 of these submissions	Missing 2 or more of these submissions